Modeling Background Linkage Disequilibrium in Human DNA

Mary Sara McPeek

Department of Statistics
Department of Human Genetics
University of Chicago

Joint work with Maoxia Zheng (Department of Statistics)

What is background linkage disequilibrium?

- Think of sampling DNA from the same region of the genome in many different individuals from a population.
- (Actually, each individual has 2 copies of each chromosome, so we get two DNA samples of a genomic region from each individual. Call each of these samples a haplotype.)
- Most sites (nucleotides) will be the same for all sampled haplotypes. The ones that differ for some haplotypes are polymorphic sites. A subset of polymorphic sites are chosen to be markers, i.e. sites at which data are collected.
- Markers we consider are binary, i.e. each haplotype can have one of two types (alleles) at a marker. For the moment say they are coded as 0 or 1.
- Sample an individual at random, and a haplotype at random from that individual. Let $Z_i \in \{0, 1\}$ be the random variable representing the allele at marker i.
- **Linkage disequilibrium (LD)** simply means that $Z_1, Z_2, \ldots$ are not independent random variables.

Importance of background LD

- Inference on population history (shared ancestry, migration, etc.)

- Mapping complex traits in humans, e.g. asthma, diabetes, hypertension
  - When a genomic region has been implicated for a trait, can use LD to narrow the region (fine-mapping).
  - Fine-mapping often involves assessing excess LD in affected individuals (cases), and comparing to background LD in controls.
  - Background LD provides a null hypothesis against which to assess excess sharing among affecteds.

$\Rightarrow$ Need to be able to model background LD.

Recombination and its role in LD

- Each individual has 2 copies of each chromosome, one maternally- and one paternally-inherited.

- Individual passes on one version of each chromosome to an offspring.

- Version passed on is a mixture of the two parental types, due to recombination.

- Recombination causes LD to break down quickly with distance along the chromosome.

- Can model segment breakpoints as a (non-homogeneous) Poisson process, which results in Markovian structure for LD.

- In addition to recombination, LD is also affected by mutation.

Empirical observations on background LD: recent developments

- Recent reports suggest relatively simple form for high-resolution LD (Daley et al. '01; Johnson et al. '01).

  - Genome divided into disjoint blocks (roughly on the order of 10^1 or 10^2 kb each), with very strong LD within blocks — each block has only a few (e.g. $\sim$2-7) commonly-occurring haplotypes.

  - Between blocks are regions over which there is lower LD (possibly recombination "hotspots" (Jeffries et al. '01)).

- NIH starting to embark on large-scale effort to identify the haplotype blocks and characterize high-resolution LD throughout the genome (HapMap project).

- Some controversy

  - Do human haplotypes have a block structure?

Long-term goals for this project

1. Rigorously assess goodness-of-fit of proposed models to available data.

- formalize models
- fit to data
- assess statistical significance of model misfit

2. Hap Map: need high-throughput method for deciding appropriate models for LD throughout entire genome.

- model selection
- parameter estimation

3. Development of mathematical (actually statistical and computational) methods that make use of Hap Map information to map complex traits in humans.

Currently working on #1, some initial progress on #3. Remainder of this talk concerns #1.

FIGURE 2

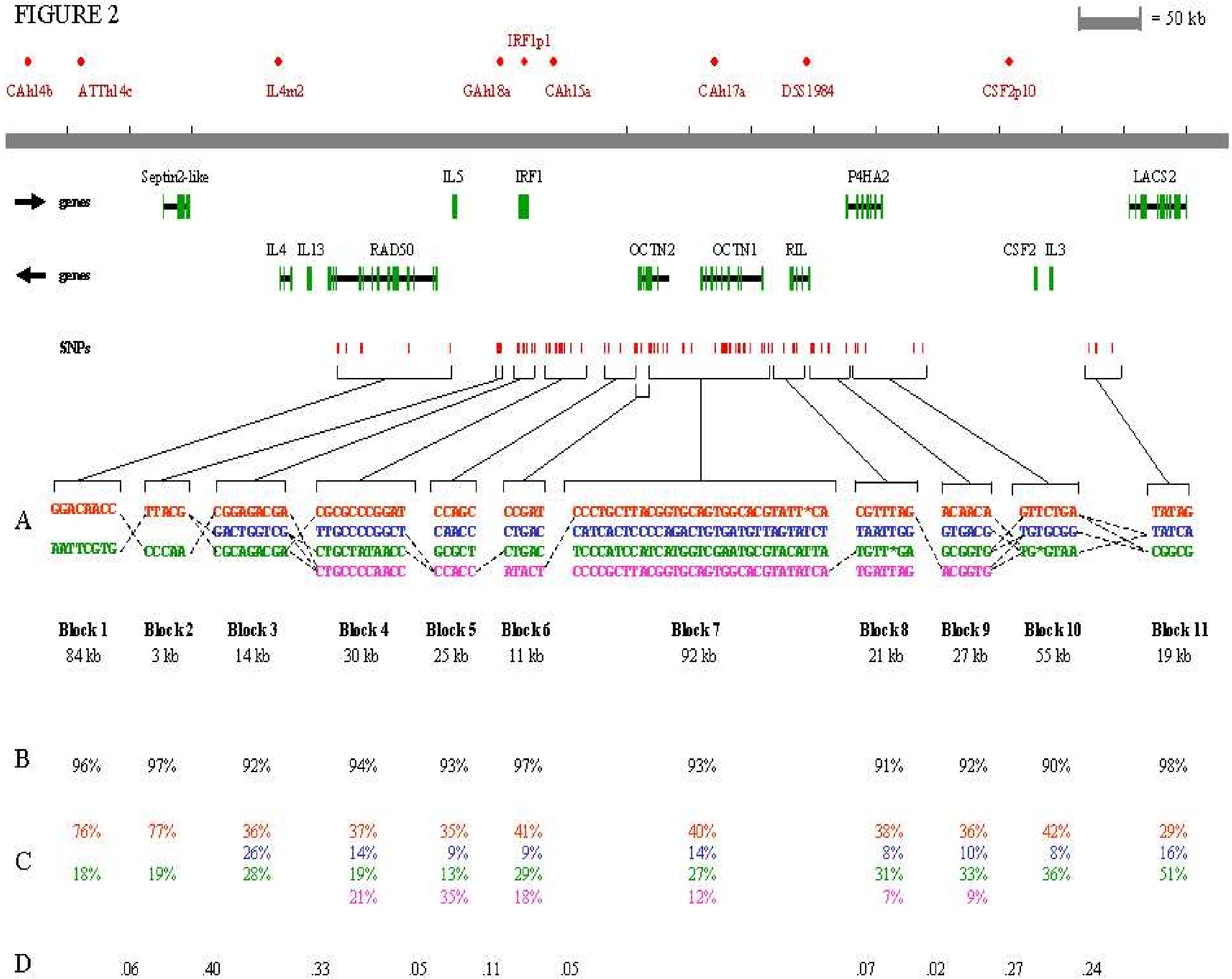
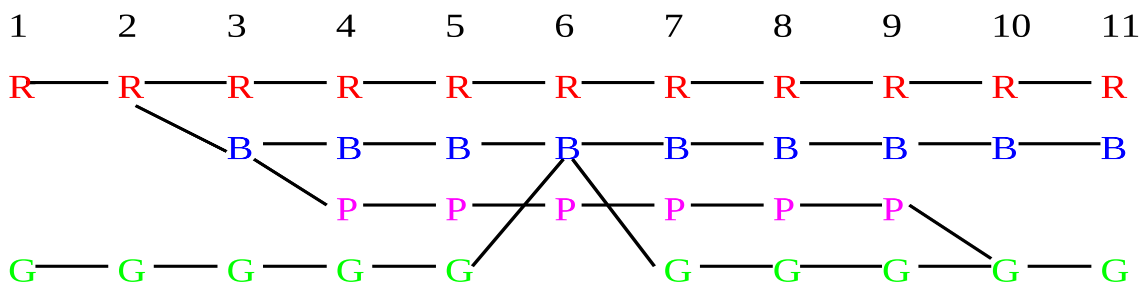
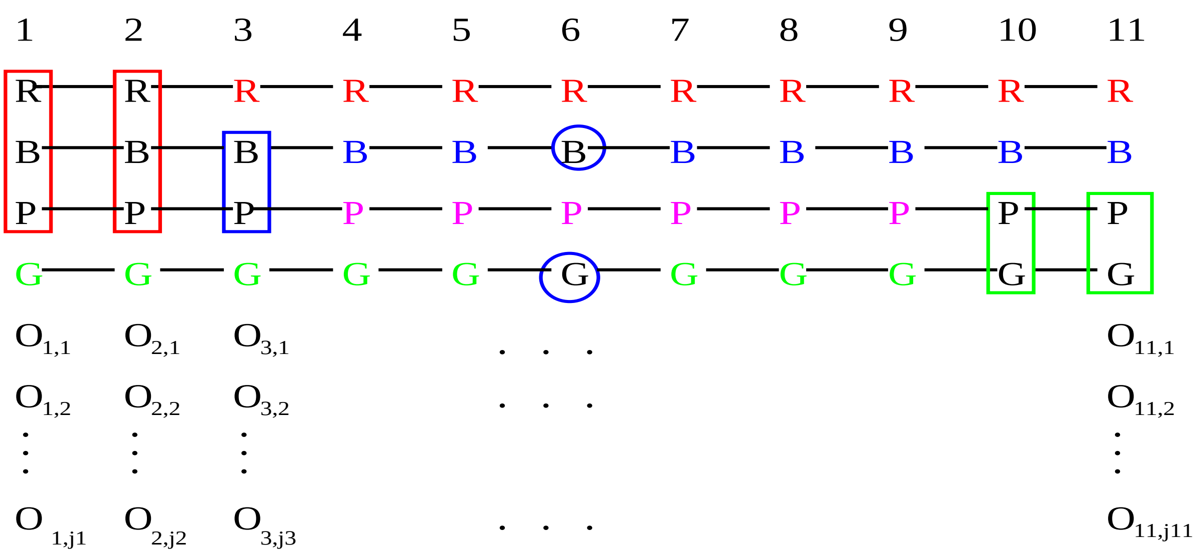


Figure 2 from Daley et al. '01.

We represent this as a Markov model. Edges represent preferred transitions.



Introduce hidden states. Also need to model the part of the data that does not fit neatly into a sequence of common haplotype blocks.



o_{ij}: “other”; uncommon words

Our model for the transition probabilities

- $P(X_{t+1} = i | X_t = i) = (1 - o)(1 - \theta_t) + \theta_t f_i$, for $i \in \{$R,B,P,G$\}$

- $P(X_{t+1} = j | X_t = i) = \theta_t f_j$, for $i \neq j$, $\{i, j\} \subset \{$R,B,P,G$\}$

- $P(X_{t+1} = o_{t+1,j} | X_t = i) = f_{o_{t+1,j}}$, for any i

- $P(X_{t+1} = j | X_t = o_{t,i}) = f_j$, for any j

- Model has property that if $P(X_1 = k) = f_k$ for all k in the state space of X_1, then $P(X_t = k) = f_k$ for all k in the state space of X_t.

- Here f_i is frequency of word i, $i \in \{$R,B,P,G$\}$ (does not depend on t).

- $f_{o_{t,j}}$ is frequency of word $o_{t,j}$.

- Require $\sum_{j=1}^{j_t} f_{o_{t,j}} = o$ for all t, and $\sum_{i \in \{\text{R,B,P,G}\}} f_i = 1 - o$.

- θ_t is "historical recombination frequency" between blocks t and $t + 1$

Data description

- Genotype, not haplotype data
  - E.g., consider block 5, and suppose an individual's 2 words in block 5 are R and G
    R $=$ C C A G C
    G $=$ G C G C T
    Actually observe the following in block 5:

nucleotide:	1	2	3	4	5
observation:	1 C, 1 G	2 C's	1 A, 1 G	1 C, 1 G	1 C, 1 T

  Based on this observation, many pairs of words are compatible with the data: e.g. (CCAGC, GCGCT) or (CCACC, GCGGT) or ... (2^4 possibilities).
  This type of missing information is called "unknown phase."

- Information at some nucleotides is missing for some individuals (very common in the data)

- Genotype data on <u>mother-father-child trios</u>, not unrelated individuals
  - four haplotypes per family
  - can reconstruct some of the phase information
  - our Markov chain for a family is actually $\{X_t\} = \{(X_t^1, X_t^2, X_t^3, X_t^4)\}$, where X_t^i is as before.
  - Missing information results in dependence among $X_t^1, X_t^2, X_t^3, X_t^4$, conditional on the data

Likelihood calculation and maximization

- We use a hidden Markov method (HMM) for likelihood calculation and maximization over the parameter values (θ_t's,f_i's,$o_{t,i}$'s).

- Consider $n = 129$ independent samples for the hidden Markov chain $\{X_t\} = \{(X_t^1, X_t^2, X_t^3, X_t^4)\}$.

- Observations $\{Y_t\}_{t=1}^L$ on each chain consist of trio genotypes for each nucleotide (when available). Conditional on $X_1, \ldots, X_L$, Y_t depends only on X_t.

- Can use forward and backward algorithms on each of the 129 families separately to get the conditional expectations of the sufficient statistics (call this vector s) for the parameters, where the sufficient statistics are linear combinations of the indicators $1(X_t = i, X_{t+1} = j)$ for each t, i, j.

- Plug conditional expectations s into complete data likelihood and maximize. Have analytical formulae for this maximization step.

- Iterate previous 2 steps.

Difficulties:

1. For a few families, the amount of information is very low, so the allowable state space is large in each block, and the algorithm is slow.

2. Many parameters to estimate relative to the amount of data.

Possible ways to address these include:

1. Use Gibbs sampling (a form of Markov chain Monte Carlo) instead of HMM to approximate the needed computations. Involves iterative conditional sampling of X_t conditional on current values of $X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_L$ and $\{Y_t\}_{t=1}^{L}$, for $t = 1, \ldots, L$.

2. Put priors on the parameters and perform Bayesian inference.

Testing goodness of fit of the model

- Is the model adequate to describe background LD?
  - Do the data look like a typical realization from the model, or do they show significant misfit?

1. Fit model $\Rightarrow$ estimate parameters (call estimated parameter vector Λ^D), calculate goodness-of-fit statistic S^D.

2. Simulate m realizations of data from model with parameters Λ^D. Keep same pattern of missing info as in real data.

3. For ith realization, estimate parameters Λ^i, calculate goodness-of-fit statistic S^i, $i = 1, \ldots, m$.

4. Find percentile rank of S^D among $S^1, \ldots, S^m$. Reject (i.e. model does not fit) if S^D is in upper α tail of distribution (e.g. $\alpha = .05$).

- Note that steps 2 and 3 involve m independent runs $\Rightarrow$ can be done in parallel

- Preliminary results

- Plans for assessing other data sets

Summary

- Initial goal of our project is to develop formal probability models that capture background LD.

- Models should be tractable and should fit well to data.

- We have developed hidden Markov models that try to capture the dependence structure in the data.

- Likelihood calculation, parameter estimation, and assessment of goodness-of-fit are computationally challenging. Current approach to the first two is maximum likelihood estimation by HMM. May also consider a Bayesian approach and Gibbs Sampling (Markov Chain Monte Carlo).

- Would like to be able to apply this approach to assess evidence for block structure of LD throughout the genome and to estimate relevant parameters.

- Our ultimate goal is to incorporate these models for background LD into mathematical (statistical and computational) methods for LD mapping of common diseases in humans.